Intégration probabiliste de sens dans la représentation de textes

Romaric Besançon, Antoine Rozenknop,
Jean-Cédric Chappelier et Martin Rajman
Laboratoire d'Intelligence Artificielle, Département Informatique
École Polytechnique Fédérale de Lausanne
e-mail: {Romaric.Besancon, Antoine.Rozenknop, Jean-Cedric.Chappelier, Martin.Rajman}@epfl.ch

Résumé - Abstract

Le sujet du présent article est l'intégration des sens portés par les mots en contexte dans une représentation vectorielle de textes, au moyen d'un modèle probabiliste. La représentation vectorielle considérée est le modèle DSIR, qui étend le modèle vectoriel (VS) standard en tenant compte à la fois des occurrences et des co-occurrences de mots dans les documents. L'intégration des sens dans cette représentation se fait à l'aide d'un modèle de Champ de Markov avec variables cachées, en utilisant une information sémantique dérivée de relations de synonymie extraites d'un dictionnaire de synonymes.
Mots-clés: Désambiguïsation, Sémantique Distributionnelle, Représentation Vectorielle, Recherche Documentaire, Champs de Markov, algorithme EM.

The present contribution focuses on the integration of word senses in a vector representation of texts, using a probabilistic model. The vector representation under consideration is the DSIR model, that extends the standard Vector Space (VS) model by taking into account both occurrences and co-occurrences of words. The integration of word senses into the co-occurrence model is done using a Markov Random Field model with hidden variables, using semantic information derived from synonymy relations extracted from a synonym dictionary.
Keywords: Word Sense Disambiguation, Distributional Semantics, Vector Space Representation, Information Retrieval, Markov Random Fields, EM algorithm.

1 Introduction

Les systèmes pour le traitement de bases d'information textuelle de grande taille reposent généralement sur la notion de similarité entre textes. Par exemple, les systèmes de Recherche Documentaire (RD) cherchent à calculer les similarités entre une requête et une collection de documents. Les techniques de classification non supervisées reposent également sur une mesure de similarité qui permet de construire des classes en regroupant des documents similaires.

Ces systèmes de similarités textuelles utilisent en général un modèle vectoriel pour la représentation des documents (Salton and Buckley, 1988; Rajman et al., 2000) pour dériver une mesure

de similarité (au sens mathématique) dans l'espace vectoriel considéré (par exemple, le cosinus de l'angle entre les vecteurs représentant les documents). Les dimensions de cet espace vectoriel sont en général associées à des unités linguistiques spécifiques, qui peuvent par exemple être des mots, des *stems* ou des lemmes. Ces unités linguistiques sont appelées *termes d'indexation*. L'idée présentée dans cet article est d'améliorer la qualité des représentations des textes en précisant le sens des unités linguistiques polysémiques. Dans ce but, une phase de désambiguïsation sémantique est intégrée dans le processus de représentation. L'information sémantique requise pour cette étape est extraite des relations de synonymie contenues dans un dictionnaire de synonymes.

Dans la section 2, nous présentons rapidement le modèle DSIR (*Distributional Semantics for Information Retrieval* – Recherche Documentaire à base de Sémantique Distributionnelle) pour la représentation de textes. Dans la section 3, nous présentons une méthode pour dériver une notion de *concepts* à partir d'un ensemble de relations de synonymie. La section 4 traite de l'intégration de ces concepts dans le modèle probabiliste, et de l'utilisation d'un algorithme de type EM pour estimer les paramètres de ce modèle. Finalement, dans la section 5, nous présentons les premiers résultats pour la validation de cette approche, appliquée à la Recherche documentaire.

2 Le modèle DSIR de représentation de textes

Les modèles vectoriels standards représentent les documents par un vecteur, appelé *profil lexical*, dont chaque composante représente l'importance dans le document du terme d'indexation associé à cette dimension. La notion d'importance (*i.e.* la composante du vecteur) se traduit habituellement par une valeur qui dépend de la fréquence du terme dans le document et, éventuellement, d'informations globales comme la fréquence en documents (Salton and Buckley, 1988; Singhal, 1997). L'idée du modèle DSIR, fondé sur la notion de *Sémantique Distributionnelle*, est d'intégrer plus d'information sémantique dans la représentation au moyen de l'utilisation de co-occurrences de termes (Rungsawang and Rajman, 1995; Rajman et al., 2000).

L'idée d'origine de la sémantique distributionnelle est que le mot prend son sens en contexte et que le sens d'un mot est donc relié à l'ensemble des contextes dans lequel il apparaît (Rajman and Bonnet, 1992). Dans l'approche DSIR, les contextes sont pris en compte au moyen des fréquences de co-occurrence. Une unité linguistique u_j est représentée par son *profil de co-occurrence* $(c_{j1}, \ldots, c_{jM})$, où M est la taille de l'ensemble des termes d'indexation et c_{ji} est la fréquence de co-occurrence entre u_j et le terme d'indexation t_i au sein d'une unité textuelle prédéfinie (fenêtre de taille fixe, phrase, paragraphe, etc). Les documents sont alors représentés par la moyenne pondérée des profils de co-occurrence des mots qu'ils contiennent, *i.e.* $d = (d_1, \ldots, d_M)$, où

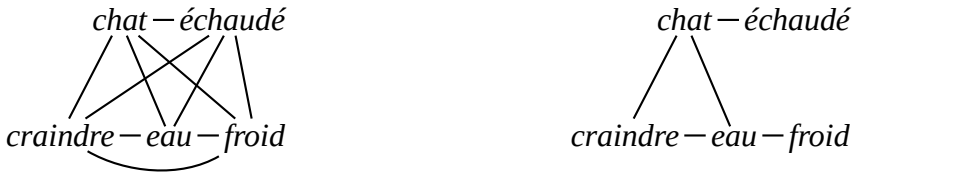
$$d_i = \sum_{u_j \in U} f_j \, c_{ji}$$

U étant l'ensemble de toutes les unités linguistiques, et f_j la fréquence de u_j dans le document d. Comme montré dans (Besançon et al., 1999; Rajman et al., 2000), le modèle DSIR a également une interprétation probabiliste.

La façon de calculer les fréquences de co-occurrence dépend des relations de co-occurrence considérées. L'approche la plus simple est de calculer toutes les co-occurrences entre toutes les unités linguistiques dans une phrase ou une fenêtre de taille donnée sur un corpus de référence.

Une approche plus fine passe par l'utilisation d'une information syntaxique supplémentaire permettant de produire des groupes syntaxiques (associés chacun à une unité linguistique particulière considérée comme tête du groupe) : on ne considére alors que les co-occurrences à l'intérieur d'un groupe syntaxique ou entre têtes de différents groupes (Besançon et al., 1999), ce qui permet d'éviter de tenir compte de co-occurrences entre unités linguistiques non syntaxiquement liées. Dans tous les cas, la phrase est transformée en un *graphe de co-occurrence* dans lequel les nœuds sont associés aux unités linguistiques considérées et les arcs représentent les relations de co-occurrence.

Prenons par exemple la phrase : "*Chat échaudé craint l'eau froide.*". Après un prétraitement permettant de ne garder que les lemmes des mots pleins[1], l'information de co-occurrence peut être calculée et les graphes de co-occurrences, correspondant à la prise en compte ou non d'un filtrage syntaxique, sont présentés dans la Figure 1.



chat échaudé craindre eau froid *(*chat échaudé) (*craindre) (*eau froid)*

(a) (b)

Figure 1: Exemples de graphes de co-occurrences , (a) sans filtrage syntaxique, (b) avec filtrage syntaxique (les étoiles indiquent les têtes des groupes syntaxiques).

L'objectif de l'intégration des sens dans le modèle de représentation des textes est de proposer un espace de représentation fondé sur les sens, c'est-à-dire un espace vectoriel dont chaque dimension est associée à un sens et non plus à un mot (terme d'indexation). Dans le cadre du modèle DSIR, cette intégration passe par le calcul des co-occurrences entre sens au lieu des co-occurrences entre mots, permettant de construire des représentations sémantiques plus riches dans les profils de co-occurrence.

Les sens qui sont associés aux dimensions de l'espace sont dérivés d'un ensemble de relations de synonymie extraites d'un dictionnaire de synonymes. La section suivante présente une méthode de construction de ces sens à partir des données brutes (et ambiguës) d'un dictionnaire de synonymes.

3 Désambiguïsation d'un ensemble de relations de synonymie

Un dictionnaire de synonymes peut être considéré comme un ensemble de mots U (entrées du dictionnaire[2]) associés chacun à un ensemble de listes de synonymes. Par exemple, quelques sens du mot *froid* et les listes de synonymes correspondantes sont (selon le dictionnaire des synonymes Hachette) :

[1] par exemple, on peut décider de ne garder que les lemmes des noms, des verbes et des adjectifs.

[2] cet ensemble correspond ici à l'ensemble d'unités linguistiques considéré dans la section 2.

- Sens 1 ⇒ *calme, de marbre, flegmatique, impassible, imperturbable, indifférent, insensible, marmoréen*;
- Sens 2 ⇒ *dédaigneux, distant, fier, hautain, réfrigérant, renfermé, réservé, sec*;
- Sens 3 ⇒ *austère, glacé, inexpressif, monotone, nu, plat, sec, sévère, simple, terne*;
- ...

Considérons alors la relation *est-synonyme-de*[3]. Comme, dans les dictionnaires standards, la transitivité n'est en général pas garantie pour cette relation, on construit tout d'abord la fermeture transitive de la relation *est-synonyme-de*, puis on extrait les sous-graphes correspondant à des composantes connexes. Par construction, une telle composante connexe contient alors un ensemble de sens recoupant une même notion. Cet ensemble de sens connexes sera appelé un *concept*.

Cependant, de façon générale, les mots apparaissant dans la liste des synonymes fournie par un dictionnaire sont souvent eux-même polysémiques et doivent être désambiguïsés pour permettre la construction des concepts : dans l'exemple précédent, le mot *sec* est synonyme de deux sens différents de *froid*. On suppose donc que ce mot a lui-même au moins deux sens différents et des approches heuristiques doivent être envisagées pour associer un sens unique à chacun des mots au sein des listes de synonymes.

Notons $w_i^1, \ldots, w_i^{n_i}$ les n_i sens d'un mot w_i, et $syn(w_i^j) \subset U$ l'ensemble de synonymes du mot w_i dans son sens j. L'heuristique utilisée ici est d'associer à chaque mot $w_i \in syn(w_k^l)$ le sens w_i^j pour lequel l'intersection entre la liste des synonymes représentant w_i^j et la liste des synonymes représentant w_k^l est maximale. Plus précisément, le sens j^* associé à w_i est défini par :

$$j^* = \underset{j}{\operatorname{argmax}} \left| syn(w_i^j) \bigcap \{w_k^l \cup syn(w_k^l) \setminus w_i^j\} \right|$$

À l'aide d'une telle heuristique, il est alors possible d'associer à chaque mot un ensemble de concepts possibles (les différents sens du mot) et à chaque concept un ensemble de mots (les réalisations du concept).

Le problème principal de l'intégration des sens dans la représentation des textes est alors de choisir les bons concepts à associer aux mots d'un corpus d'apprentissage, pour le calcul des co-occurrences entre concepts.

4 Intégration des sens dans le modèle DSIR

L'approche choisie pour la désambiguïsation sémantique des mots est de suivre l'intuition qui est à la base de la Sémantique Distributionnelle : on fait l'hypothèse que la séquence de concepts à associer à une séquence de mots est celle qui maximise la probabilité d'affecter des concepts aux nœuds du graphe de co-occurrence dérivé de la séquence de mots. En d'autres termes, pour une configuration de mots w associée aux nœuds d'un graphe de co-occurrence g, la configuration de concepts associée c^* est telle que

$$c^* = \underset{c}{\operatorname{argmax}}\, p(c|w,g) = \underset{c}{\operatorname{argmax}}\, p(w|c,g)\; p(c|g)$$

[3] on suppose que cette relation est symétrique, ce qui correspond à l'intuition.

On notera, pour tout graphe g, N_g l'ensemble de ses nœuds, A_g l'ensemble de ses arcs, et $w^1, \ldots, w^{|N_g|}$ (resp. $c^1, \ldots, c^{|N_g|}$) une configuration de mots (resp. de concepts) associée aux nœuds du graphe.

Pour calculer $p(w|c,g)\ p(c|g)$, on fait les hypothèses suivantes :

- **H1:** le conditionnement imposé à l'association d'un mot à un nœud est limité au seul concept associé à ce nœud, ce qui se traduit par : $p(w|c,g) = \prod_i p(w^i|c^i)$;
- **H2:** le conditionnement de l'association d'un concept à un nœud du graphe est limité aux concepts associés aux nœuds voisins : $p(c^i|(c_j)_{j\in N_g}) = p(c^i|(c^j)_{j\in \mathcal{V}_i})$, où $\mathcal{V}_i$ est le voisinage du nœud i dans le graphe g.

L'hypotyhèse **(H2)** de conditionnement limité au voisinage, induit une structure de Champ de Markov (*Markov Random Field* – MRF) pour le modèle probabiliste. D'après le théorème de Hammerlsey (Besag, 1974), la distribution de probabilité $p(c|g)$ est alors une distribution de Gibbs, c'est-à-dire une distribution de la forme :

$$p(c|g) = \frac{1}{Z_0} \exp \sum_{\gamma\in\Gamma} V_\gamma(c)$$

où Z_0 est un facteur de normalisation ($Z_0 = \sum_c \exp[\sum_{\gamma\in\Gamma} V_\gamma(c)]$), Γ est un ensemble de cliques dans g et $V_\gamma(c)$ est une fonction, appelée *potentiel*, qui associe une valeur réelle à chacune de ces cliques. Comme, dans notre cas, on ne regarde que des relations de co-occurrence, les cliques considérées sont d'ordre au plus 2, et on peut donc réécrire $p(c|g)$ comme:

$$p(c|g) = \frac{1}{Z_0} \exp \left[\sum_{i\in N_g} V(c^i) + \sum_{(i,j)\in A_g} V(c^i, c^j) \right]$$

La probabilité à maximiser pour obtenir la configuration de concepts optimale c^* est donc :

$$p(w|c,g)\ p(c|g) = \frac{1}{Z} \exp \left[\sum_{i\in N_g} V_e(w^i, c^i) + \sum_{i\in N_g} V(c^i) + \sum_{(i,j)\in A_g} V(c^i, c^j) \right]$$

où $V_e(w^i, c^i)$ est le potentiel associé à la probabilité d'émission $p(w^i|c^i)$.

Cette distribution de probabilité correspond à un modèle paramétrique, dont l'ensemble des paramètres est $\theta = \{(V(c_i))_{c_i\in\mathcal{C}}, (V(c_i,c_j))_{c_i,c_j\in\mathcal{C}}, (V_e(w_i,c_j))_{w_i\in U, c_j\in\mathcal{C}}\}$, où $\mathcal{C}$ est l'ensemble des concepts. Pour trouver des valeurs pertinentes pour ces paramètres, on utilise l'approche standard qui consiste à choisir les valeurs qui maximisent la log-vraisemblance du corpus d'apprentissage C (les graphes de co-occurrence g étant donnés par une analyse syntaxique préalable des phrases de C).

On utilise dans ce but un algorithme dérivé des algorithmes *Estimation-Maximisation (EM)* (Dempster et al., 1977) et *Improved Iterative Scaling (IIS)* (Lafferty, 1996; Pietra et al., 1997), dont l'objectif est de déterminer l'ensemble de paramètres θ qui maximise l'espérance de la log-vraisemblance $L_\theta(C) = \sum_{(w,g)\in C} \log p_\theta(w|g)$.

La structure générale de cette approche itérative est la suivante : étant donné un ensemble de paramètres θ, on cherche l'ensemble de paramètre θ' qui maximise l'espérance de la log-vraisemblance sachant les paramètres θ. Cela revient à maximiser la fonction $Q(\theta, \theta')$ suivante :

$$Q(\theta, \theta') = \sum_{(w,g) \in C} \sum_{c \in \mathcal{C}_g(w)} p_\theta(c|w,g) \log \frac{p_{\theta'}(w,c|g)}{p_\theta(w,c|g)}$$

où $\mathcal{C}_g(w)$ est l'ensemble des configurations de concepts possibles sur le graphe g, sachant les mots associés aux nœuds du graphe.

Dans le cas d'une distribution de Gibbs, la log-vraisemblance est difficile à calculer, en raison du caractère global de la constante de normalisation Z, trop coûteuse à évaluer. L'approche généralement utilisée (en particulier dans le domaine du traitement d'image) pour résoudre ce problème est de remplacer la vraisemblance par une pseudo-vraisemblance (Besag, 1974; Chalmond, 1989), définie comme le produit des probabilités conditionnelles :

$$\mathrm{PL}_\theta(c|g) = \prod_{i \in N_g} p(c^i | (c^j)_{j \in \mathcal{V}_i})$$

Il a été montré que l'estimateur ainsi obtenu est aussi consistant que l'estimateur du maximum de vraisemblance (Gidas, 1988).

Pour ce qui est de la distribution jointe des variables observées et cachées, la pseudo-vraisemblance est définie par (Chalmond, 1989) :

$$\mathrm{PL}_\theta(w,c|g) = p_\theta(w|c)\mathrm{PL}_\theta(c|g)$$

et on cherche alors à maximiser la fonction $Q(\theta, \theta')$ suivante :

$$Q(\theta, \theta') = \sum_{(w,g) \in C} \sum_{c \in \mathcal{C}_g(w)} p_\theta(c|w,g) \log \frac{\mathrm{PL}_{\theta'}(w,c|g)}{\mathrm{PL}_\theta(w,c|g)} \quad (1)$$

Comme il n'est en général pas possible de maximiser directement cette expression, l'approche usuelle est d'utiliser une fonction auxiliaire $A(\theta, \theta')$ qui constitue une borne inférieure pour la fonction $Q(\theta, \theta')$[4].

En effectuant les calculs à partir de la définition (1), on peut séparer la fonction Q en deux fonctions que l'on pourra maximiser séparément :

$$Q(\theta, \theta') = Q_E(\theta, \theta') + Q_C(\theta, \theta')$$

$$\text{avec } \begin{cases} Q_E(\theta, \theta') = \sum_{(w,g) \in C} \sum_{c \in \mathcal{C}_g(w)} p_\theta(c|w,g) \log \dfrac{p_{\theta'}(w|c)}{p_\theta(w|c)} \\ Q_C(\theta, \theta') = \sum_{(w,g) \in C} \sum_{c \in \mathcal{C}_g(w)} p_\theta(c|w,g) \log \dfrac{\mathrm{PL}_{\theta'}(c|g)}{\mathrm{PL}_\theta(c|g)} \end{cases}$$

[4] maximiser $A(\theta, \theta')$, si ce maximum est strictement positif, assure que les paramètres θ' définissent un meilleur modèle que θ.

Et on dérive ainsi les fonctions auxiliaires suivantes (en explicitant les valeurs de $p_\theta(w|c)$ et $\mathrm{PL}_\theta(c|g)$) :

$$A_E(\theta,\theta') = \sum_{(w,g)} \sum_{c\in\mathcal{C}_g(w)} p_\theta(c|w,g) \sum_{i\in N_g} \left[\Delta V_e(w^i,c^i) - \sum_{w\in c^i} p_\theta(w|c^i) \exp \Delta V_e(w,c^i) \right]$$

$$A_C(\theta,\theta') = \sum_{(w,g)} \sum_{c\in\mathcal{C}_g(w)} p_\theta(c|w,g) \sum_{i\in N_g} \left[\Delta V(c^i) + \sum_{j\in\mathcal{V}_i} \Delta V(c^i,c^j) \right.$$
$$\left. - \sum_{c_k\in\mathcal{C}} p_\theta(c_k|(c^j)_{j\in\mathcal{V}_i}) \frac{1}{\lambda_i} \left(\sum_{j\in\mathcal{V}_i} \exp \lambda_i \, \Delta V(c_k,c^j) + \exp \lambda_i \, \Delta V(c_k) \right) \right]$$

où les ΔV sont les différences de potentiel entre les modèles θ' et θ, et λ_i est le nombre de voisins du nœud i augmenté de 1 ($\lambda_i = |\mathcal{V}_i| + 1$).

Les dérivées $\frac{\partial A_E(\theta,\theta')}{\partial \Delta V_e(w,c)}$ ne dépendent alors que des $\Delta V_e(w,c)$. On peut donc trouver les paramètres $V_e(w,c)$ en calculant itérativement les valeurs $\Delta V_e(w,c)$ qui maximisent $A_E(\theta,\theta')$ (en utilisant des méthodes comme l'algorithme de Newton pour résoudre les équations $\frac{\partial A_E(\theta,\theta')}{\partial \Delta V_e(w,c)} = 0$) et remplacer $V_e(w,c)$ par $V_e(w,c) + \Delta V_e(w,c)$ jusqu'à convergence. On fait de même pour les paramètres $\Delta V(c)$ et $\Delta V(c_i,c_j)$.

Une fois les paramètres optimaux obtenus, on peut, pour chacune des phrases du corpus, calculer la configuration de concepts la plus probable qui lui est associée. On obtient donc un corpus composé de graphes de co-occurrences dont les nœuds sont associés à des concepts, à partir duquel on peut calculer les co-occurrences nécessaires à la représentation des textes dans l'espace des concepts.

5 Évaluation

L'approche présentée dans cet article a été évaluée pour une tâche de Recherche Documentaire, en utilisant un corpus d'articles du journal La Monde, provenant de la campagne d'évaluation AMARYLLIS. Cette évaluation n'est pas directement centrée sur les résultats obtenus pour la désambiguïsation sémantique (c'est-à-dire l'affectation des concepts les plus probables à une phrase), mais plutôt sur l'impact de l'utilisation des concepts produits pour une application du type Recherche Documentaire. Pour cette tâche, les documents de la base documentaire et les requêtes sont représentés dans l'espace vectoriel des concepts, à l'aide d'un modèle DSIR utilisant les co-occurrences entre concepts au lieu des co-occurrences entre mots.

Le travail de désambiguïsation d'une liste de synonymes dérivée d'un dictionnaire de synonymes a été mise en œuvre séparément et un ensemble de concepts en ont été dérivées (Pfister, 2000).

Pour la tâche de Recherche Documentaire, un lexique de 7073 mots (lemmes) est utilisé, qui correspond à un total de 2936 concepts (le nombre moyen de mots par concepts est 2.7). Le corpus de référence est composé de 9574 documents et de 4 requêtes. Les résultats, en termes de précision/rappel, présentés en Figure 2 montrent une amélioration des performances lors de l'utilisation de la représentation en concepts des documents et des requêtes (courbe "*concepts*") par rapport à la représentation en mots (courbe "*mots*"). Un test supplémentaire a été effectué en associant le même concept à toutes les occurrences d'un mot donné, par exemple, en

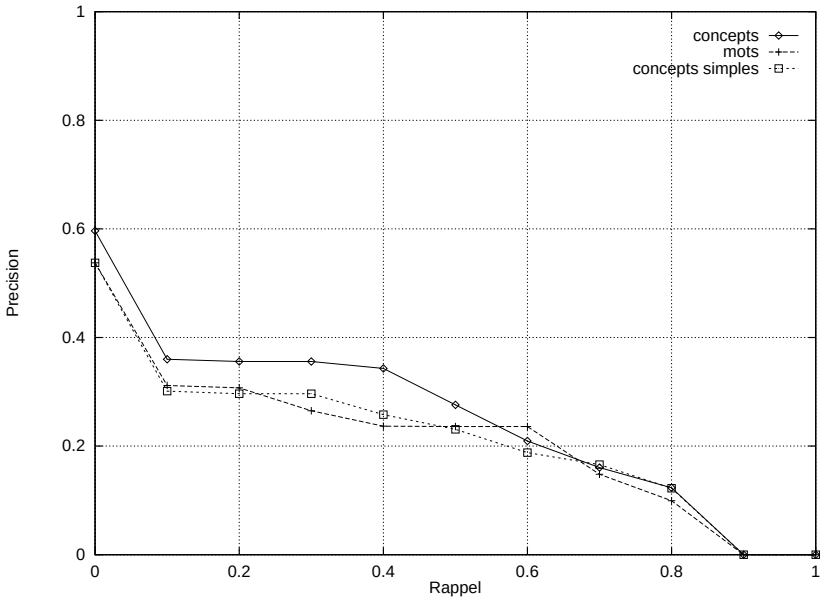


Figure 2: Résultats de l'intégration des sens pour une tâche de Recherche Documentaire sur le corpus "Le Monde".

prenant toujours le premier de la liste des concepts possibles du mot. Les résultats de ce test (représentés par la courbe "*concepts simples*") montrent que cette représentation n'améliore pas les performances par rapport à la représentation directe en mots, ce qui semble confirmer que l'amélioration des performances est effectivement due à la pertinence de l'association des concepts aux mots par la méthode présentée. D'autres expériences de validation sont néanmoins nécessaires pour fournir plus de détails sur les caractéristiques de la méthode qui expliquent cette amélioration.

6 Conclusion

Nous présentons dans cet article un modèle de représentation à base de Champ de Markov permettant d'intégrer les sens des mots dans un modèle probabiliste de représentation de textes. Les champs de Markov semblent fournir un bon cadre pour la représentation de ce type d'information de voisinage non-orienté, et peuvent également être envisagés pour la modélisation directe des co-occurrences de mots dans la représentation de documents. Les premiers résultats pour l'évaluation de cette représentation pour une tâche de recherche documentaire montrent une amélioration des performances, mais une évaluation plus poussée devrait être mise en œuvre. D'autre part, la disponibilité d'un modèle de co-occurrence de concepts permet également de mettre en œuvre des techniques de désambiguïsation sémantique, et ce modèle devrait aussi être évalué précisément pour cette seule tâche, par une méthode adaptée, et pourrait être comparé à d'autres techniques de désambiguïsation.

Références

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of Royal Statistics Society*, 36:192–236.

Besançon, R., Rajman, M., and Chappelier, J.-C. (1999). Textual similarities based on a distributional approach. In *International Workshop on Similarity Search (IWOSS99)*, Florence, Italy.

Chalmond, B. (1989). An iterative gibbsian technique for reconstruction of m-ary images. *Pattern Recognition*, 22(6):747–761.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Jounal of Royal Statistics Society*, 39:185–197.

Gidas, B. (1988). Consistency of maximum likelihood and pseudolikelihood estimators for gibbs distributions. In Fleming, W. and Lions, P., editors, *Stochastic Differential System, Stochastic Control Theory and Applications*, pages 129–145. Springer, New York.

Lafferty, J. (1996). Gibbs-markov models. *Computing Science and Statistics*, 27:370–377.

Pfister, J.-P. (2000). Désambiguisation d'un dictionnaire de synonymes. Technical report, EPFL.

Pietra, S., Pietra, V., and Lafferty, J. (1997). Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4):380–393.

Rajman, M., Besançon, R., and Chappelier, J.-C. (2000). Le modèle DSIR : Une approche à base de sémantique distributionnelle pour la recherche documentaire. *Traitement Automatique des Langues*, 41(2).

Rajman, M. and Bonnet, A. (1992). Corpora-base linguistics: new tools for natural language processing. In *1st Annual Conference of the Association for Global Strategic Information*, Bad Kreuznach, Germany.

Rungsawang, A. and Rajman, M. (1995). Textual information retrieval based on the concept of distributional semantics. In *proc. of JADT'95 (3rd International Conference on Statistical Analysis of Textual Data)*, Rome.

Salton, G. and Buckley, C. (1988). Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 24:513–523.

Singhal, A. (1997). *Term Weighting Revisited*. PhD thesis, Department of Computer Science, Cornell University.